Home Assignment - Backend Engineer Data Group

Hello again and thanks for taking the time to complete this modest home assignment. All in all we estimate that it should take about 6 hours to complete. The assignment is to be completed in Nodejs or Python, and please feel free to reach out to us with any questions whatsoever.

THE ASSIGNMENT

The data group is responsible for consuming the data from all around the web, analyzing it, and storing it for later use by researchers and other processes. Our OSINT analysts found some wikipedia entries that contain malicious content that could be relevant to our offerings. Create a program that can scrape any wikipedia entry and model the data in a useful accessible format.

The program should recieve the following params:

1. A link to a wikipedia entry
2. The depth to which we want to crawl from it
3. The crawling algorithm to use (bfs/dfs)

It will then start scraping the entry and crawling from it into any other wikipedia entries it finds according to the algorithm specified. For each entry the program should extract and model the data of the entry into a useful database of entries and the relations between them as well as other useful info that can be extracted from an entry such as links.

Bonus: Design the program to be able to run as a horizontally scaleable cluster where any specific crawl can be distributed between multiple processes.

Package this all up into a private github repo with an easy to understand readme and share it with zivk@activefence.com and org@activefence.com. Make sure to include this pdf in the repo as well.

Very best of luck, we're sure you'll do great!
Ziv: 052-6555995, Or: 050-2272550